

Singular value decomposition

Victor Kitov

Definitions

Consider matrix $X \in \mathbb{R}^{N \times D}$. For this matrix:

- eigenvalues of matrix $X^T X$ are called *singular values*.
- orthonormal eigenvectors of $X^T X$ are called *right singular vectors*.
- orthonormal eigenvectors of XX^T are called *left singular vectors*.

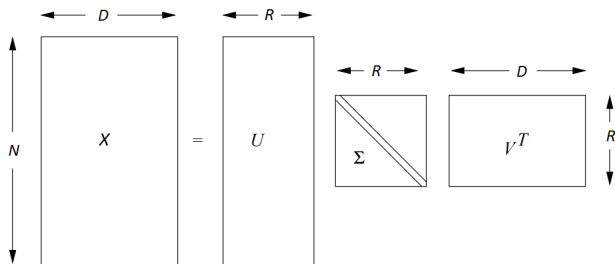
Principal component a_i is the i -th right singular vector of X , corresponding to i -th largest singular value λ_i .

SVD decomposition

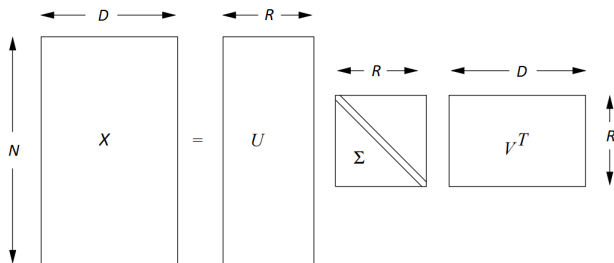
Every matrix $X \in \mathbb{R}^{N \times D}$, $\text{rank } X = R$, can be decomposed into the product of three matrices:

$$X = U \Sigma V^T$$

where $U \in \mathbb{R}^{N \times R}$, $\Sigma \in \mathbb{R}^{R \times R}$, $V^T \in \mathbb{R}^{R \times D}$, and $\Sigma = \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_R\}$, $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_R \geq 0$, $U^T U = I$, $V^T V = I$. $I \in \mathbb{R}^{D \times D}$ denotes identity matrix.



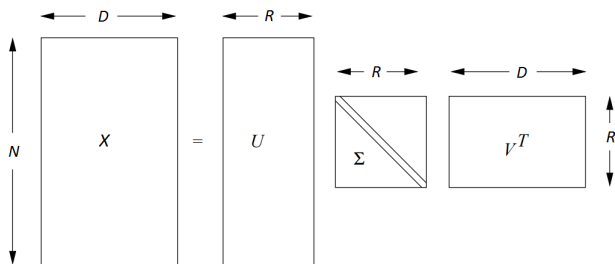
Interpretation of SVD



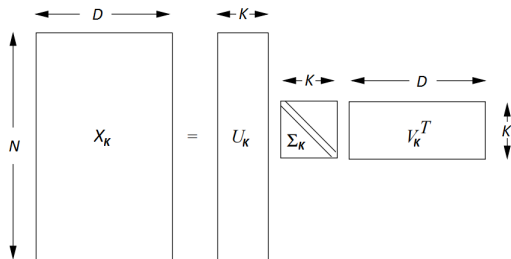
For X_{ij} let i denote objects and j denote properties.

- U represents standardized coordinates of concepts
- V^T represents standardized concepts representations
- Σ shows the magnitudes of presence of standardized concepts in X .

Original SVD decomposition



Reduced SVD decomposition



Simplification to rank $K \leq R$:

$$X_K = U_K \Sigma_K V_K$$

$$\Sigma = \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_K, \sigma_{K+1}, \dots, \sigma_R\} \longrightarrow \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_K\} = \Sigma_K$$

$$U = [u_1, u_2, \dots, u_K, u_{K+1}, \dots, u_R] \longrightarrow [u_1, u_2, \dots, u_K] = U_K$$

$$V = [v_1, v_2, \dots, v_K, v_{K+1}, \dots, v_R] \longrightarrow [v_1, v_2, \dots, v_K] = V_K$$

Properties of reduced SVD decomposition

- Suppose $X \in \mathbb{R}^{N \times D}$, $\text{rank } X = R$, is approximated with $X_K = U_K \Sigma_K V_K$. Then:
 - $\text{rank } X_K = K$.
 - $X_K = \arg \min_{B: \text{rank } B \leq K} \|X - B\|$
- Which K to choose?
 - Define Frobenius norm $\|X\|_F^2 = \sum_{n=1}^N \sum_{d=1}^D x_{nd}^2$
 - $\|X\|_F^2 = \sum_{i=1}^R \sigma_i^2$
 - $\|X_K\|_F^2 = \sum_{i=1}^K \sigma_i^2$
 - Choose $K = \arg \min_K \left\{ \frac{\|X_K\|_F^2}{\|X\|_F^2} \geq t \right\}$, where t is some threshold, say $t = 0.95$.

Table of Contents

- 1 Applications of SVD
- 2 Recommendation system with SVD

Memory efficiency

Storage costs of $X \in \mathbb{R}^{N \times D}$, assuming $N \geq D$ and each element taking 1 byte:

Memory storage costs

representation of X	memory requirements
original X	$O(ND) = O(\min\{N, D\} \max\{N, D\})$
fully SVD decomposed	$NR + R^2 + RD = O(R \max\{N, D\})$
reduced SVD to rank K	$NK + K^2 + KD = O(K \max\{N, D\})$

Performance efficiency

Suppose we have N documents, vocabulary size is D , typically $D \geq N$.

- $X \in \mathbb{R}^{N \times D}$ represents normalized vector representation of documents
- $q \in \mathbb{R}^D$ represents normalized vector representation of search query
-

$$X \approx X_K = \underbrace{U_K \Sigma_K V_K^T}_B = B V_K^T, \quad B \in \mathbb{R}^{N \times K}$$

Document x_i relevance is proportional to $\langle x_i, q \rangle$, so to find matching documents we need to calculate

$$Xq = [\langle x_1, q \rangle, \dots, \langle x_N, q \rangle]^T.$$

Direct multiplication Xq takes

$O(ND) = O(\max\{N, D\} \min\{N, D\})$ operations.

$X_K q = U_K \Sigma_K V_K^T q = B V_K^T q$. $V_K^T q$ takes $O(DK)$ multiplications and $B V_K^T q$ takes $O(NK)$, so total complexity is $O(K \max\{N, D\})$.

SVD for square non-degenerate matrix

For square non-degenerate matrix X :

- $X \in \mathbb{R}^{D \times D}$, $\text{rg } X = D$, so $U \in \mathbb{R}^{D \times D}$, $V \in \mathbb{R}^{D \times D}$, $U^{-1} = U^T$, $V^{-1} = V^T$.
- U , V^T represent rotations, Σ represents scaling, every square matrix may be represented as superposition of rotation, scaling and another rotation.
- For full rank X :

$$X^{-1} = V \Sigma^{-1} U^T,$$

$$\text{since } XX^{-1} = U \Sigma V^T V \Sigma^{-1} U^T = I.$$

Table of Contents

- 1 Applications of SVD
- 2 Recommendation system with SVD

Example

	Terminator	Gladiator	Rambo	Titanic	Love story	A walk to remember
Andrew	4	5	5	0	0	0
John	4	4	5	0	0	0
Matthew	5	5	4	0	0	0
Anna	0	0	0	5	5	5
Maria	0	0	0	5	5	4
Jessika	0	0	0	4	5	4

Example

$$U = \begin{pmatrix} 0. & 0.6 & -0.3 & 0. & 0. & -0.8 \\ 0. & 0.5 & -0.5 & 0. & 0. & 0.6 \\ 0. & 0.6 & 0.8 & 0. & 0. & 0.2 \\ 0.6 & 0. & 0. & -0.8 & -0.2 & 0. \\ 0.6 & 0. & 0. & 0.2 & 0.8 & 0. \\ 0.5 & 0. & 0. & 0.6 & -0.6 & 0. \end{pmatrix}$$

$$\Sigma = \text{diag}\{(14. \quad 13.7 \quad 1.2 \quad 0.6 \quad 0.6 \quad 0.5)\}$$

$$V^T = \begin{pmatrix} 0. & 0. & 0. & 0.6 & 0.6 & 0.5 \\ 0.5 & 0.6 & 0.6 & 0. & 0. & 0. \\ 0.5 & 0.3 & -0.8 & 0. & 0. & 0. \\ 0. & 0. & 0. & -0.2 & 0.8 & -0.6 \\ -0. & -0. & -0. & 0.8 & -0.2 & -0.6 \\ 0.6 & -0.8 & 0.2 & 0. & 0. & 0. \end{pmatrix}$$

Example (excluded insignificant concepts)

$$U_2 = \begin{pmatrix} 0. & 0.6 \\ 0. & 0.5 \\ 0. & 0.6 \\ 0.6 & 0. \\ 0.6 & 0. \\ 0.5 & 0. \end{pmatrix}$$

$$\Sigma_2 = \text{diag}\{(14. \quad 13.7)\}$$

$$V_2^T = \begin{pmatrix} 0. & 0. & 0. & 0.6 & 0.6 & 0.5 \\ 0.5 & 0.6 & 0.6 & 0. & 0. & 0. \end{pmatrix}$$

Concepts may be

- patterns among movies (along j) - action movie / romantic movie
- patterns among people (along i) - boys / girls

Dimensionality reduction case: patterns along j axis.

Applications

- Example: new movie rating by new person

$$x = (5 \ 0 \ 0 \ 0 \ 0 \ 0)$$

- **Dimensionality reduction:** map x into concept space:

$$y = V_2^T x = (0 \ 2.7)$$

- **Recommendation system:** map y back to original movies space:

$$\hat{x} = yV_2^T = (1.5 \ 1.6 \ 1.6 \ 0 \ 0 \ 0)$$

Frobenius norm

- Frobenius norm of matrix X is $\|X\|_F \stackrel{df}{=} \sqrt{\sum_{n=1}^N \sum_{d=1}^D x_{nd}^2}$
- Using properties $\|X\|_F^2 = \text{tr} XX^T$ and $\text{tr} AB = \text{tr} BA$, we obtain:

$$\begin{aligned}\|X\|_F^2 &= \text{tr}[U\Sigma V^T V\Sigma U^T] = \text{tr}[U\Sigma^2 U^T] = \\ &= \text{tr}[\Sigma^2 U^T U] = \text{tr}[\Sigma^2] = \sum_{r=1}^R \sigma_r^2\end{aligned}\quad (1)$$

Matrix approximation

Consider approximation $X_k = U\Sigma_k V^T$, where $\Sigma_k = \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_k, 0, 0, \dots, 0\} \in \mathbb{R}^{R \times R}$.

Theorem 1

X_k is the best approximation of X retaining k concepts.

Proof: consider matrix $Y_k = U\Sigma' V^T$, where Σ' is equal to Σ except some $R - k$ elements set to zero:

$\sigma'_{i_1} = \sigma'_{i_2} = \dots = \sigma'_{i_{R-k}} = 0$. Then, using (1)

$$\|X - Y_k\|_F^2 = \|U(\Sigma - \Sigma')V^T\|_F^2 = \sum_{p=1}^{R-k} \sigma_{i_p}^2 \leq \sum_{p=1}^{R-k} \sigma_p^2 = \|X - X_k\|_F^2$$

since $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_R \geq 0$.

Matrix approximation

How many components to retain?

General case: Since

$$\|X - X_k\|_F^2 = \|U(\Sigma - \Sigma_k)V^T\|_F^2 = \sum_{i=k+1}^R \sigma_i^2$$

a reasonable choice is k^* such that

$$\frac{\|X - X_{k^*}\|_F^2}{\|X\|_F^2} = \frac{\sum_{i=k^*+1}^R \sigma_i^2}{\sum_{i=1}^R \sigma_i^2} \geq \text{threshold}$$

Visualization: 2 or 3 components.

Theorem 2

For any matrix Y_k with $\text{rank } Y_k = k$: $\|X - X_k\|_F \leq \|X - Y_k\|_F$

Finding U and V

- **Finding V**

$X^T X = (U \Sigma V^T)^T U \Sigma V^T = (V \Sigma U^T) U \Sigma V^T = V \Sigma^2 V^T$. It follows that

$$X^T X V = V \Sigma^2 V^T V = V \Sigma^2$$

So V consists of eigenvectors of $X^T X$ with corresponding eigenvalues $\sigma_1^2, \sigma_2^2, \dots, \sigma_R^2$.

- **Finding U :**

$XX^T = U \Sigma V^T (U \Sigma V^T)^T = U \Sigma V^T V \Sigma U^T = U \Sigma^2 U^T$. So

$$XX^T U = U \Sigma^2 U^T U = U \Sigma^2.$$

So U consists of eigenvectors of XX^T with corresponding eigenvalues $\sigma_1^2, \sigma_2^2, \dots, \sigma_R^2$.

Comments

- Denote the average $\bar{X} \in \mathbb{R}^D$: $\bar{X}_j = \sum_{i=1}^N x_{ij}$
- Denote the n-th row of X be $X_n \in \mathbb{R}^D$: $X_{nj} = x_{nj}$
- For centered X sample covariance matrix $\hat{\Sigma}$ equals:

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{N} \sum_{n=1}^N (X_n - \bar{X})(X_n - \bar{X})^T = \frac{1}{N} \sum_{n=1}^N X_n X_n^T \\ &= \frac{1}{N} X^T X\end{aligned}$$

- V consists of **principal components** since
 - V consists of eigenvectors of $X^T X$,
 - principal components are eigenvectors of $\hat{\Sigma}$ and
 - $\hat{\Sigma} \propto X^T X$.